CHAPTER 9
EQUATING AND LINKING

INTRODUCTION

To obtain meaningful results across multiple years of the KIRIS Assessment, it is necessary that the scores from one year mean the same thing (i.e., can be interpreted the same way) as scores from previous years. In an accountability system that depends upon comparing growth from one year to the next, measurement of such growth would not be possible if the scores from previous years had a different meaning than the scores for the current year. To obtain scores that mean the same thing across years one must, in psychometric terms, equate. Year to year equating places scores obtained in different years on the same scale of measurement.

If the assessment for each year has multiple forms, like KIRIS, a within year equating must also be performed. Not only must the year-to-year scale measurement be the same, but also the within year score measurement must be the same; no mater what form of the test was taken. For a given student ability, that student must be afforded the opportunity to achieve the same equated score on any form of KIRIS he or she was randomly given.

Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. “Equating adjusts for differences in different forms that are built to be similar in difficulty and content" (Kolen and Brennan, 1995). This need for equating is necessary because of various changes in the test items used in each particular testing year. This within year need along with the across year testing grade shift in some of the content areas requires equating. Because of a testing shift to adjacent grades of several content areas equating must be done so as to adjust data to compensate for that shift. Below is a list of areas where equating was necessary.

1. Within-year forms equating. Each of the 12 forms of a given content area contained two unique matrix sampled items. These unique matrix sampled items cause each form to be not identical. Forms of a test should be interchangeable.
2. Across-year test equating. Each year of KIRIS assessment uses different items to measure content competency. Each year’s competency must be adjusted for test difficulty differences between years. This assessment adjustment will place the current years test on the same scale as the 12 forms from the previous year’s assessment.
3. Grade Shift adjustments due to change in grades tested. In 1995 and 1996 KIRIS testing was conducted at grades 4, 8, and 11. In 1997 and 1998 KIRIS elementary school testing was split between 4th and 5th grades and middle school testing was split between 7th and 8th grades. Ability level changes, caused only

because the assessment grade level shifts, need to be adjusted to compensate for this grade level shift.

The remainder of this chapter discusses the various equating methods used to respond to: 1) Within-year forms equating, 2) Across-year test equating, and 3) Grade Shift adjustments due do change in grades tested.

WITHIN-YEAR FORMS EQUATING

Reading, Mathematics, Science and Social Studies

INTRODUCTION

This section provides a summary of the results of equating the reading, mathematics, science, and social studies open-response portions of the KIRIS assessments in Accountability Cycle 3 (years 1995, 1996, 1997, and 1998). It should be noted that in 1997 and 1998 KIRIS assessments also included multiple-choice questions for each content area. However, these multiple-choice questions were not part of Cycle 3 school accountability. This chapter will only focus on the open-response items of KIRIS.

Test Structure Each content area had 12 forms of the test. Each form contained a set of items common to all forms (Reading, Mathematics, Science, and Social Studies) and a set of matrix sampled items unique to the 12 forms of that content area. There were five common items at grades 4 and 5 and six common items at grades 7, 8, and 11 in 1995 and 1996. In 1997 and 1998, to reduce testing time, each form was shortened and provided four common items for each form, at all grade and content areas. However, Arts & Humanities and Practical Living/Vocational Studies had no common items during Cycle 3. In all four years of Cycle 3, there were two operational matrix sampled items administered in each of the 12 forms in each of the content area (see Tables 3-1a though 3-3b for additional information) and one matrixed pretest item administered in each of the 12 forms.

Equating Design. There are several general equating designs in use; Random groups, Single group with counterbalancing, and Common – item nonequivalent groups (Kolen and Brennan, 1995). The KIRIS assessment uses the common – item nonequivalent groups design. This design is appropriate because the KIRIS assessment uses multiple forms per testing year.

In the common – item nonequivalent groups design each form of the test must have a set of items in common. As noted above the KIRIS test has between 4 and 6 open response items in common between each of the 12 forms in each of the content areas (excluding arts & humanities and practical living & vocational studies). The common items within each year/grade/content area were in the exact same location and worded exactly the same in each of the 12 forms. This procedure is especially important in light of the National Assessment of Educational Progress (NAEP) reading anomaly found in 1986. A large drop in reading results (from 1984 to 1986) for both 17 and 9 year olds was attributed to differences in context in which the common items appeared (Zwick, 1991). Additionally, to insure a random distribution of the 12 forms of the KIRIS assessment the forms were spiraled. The packaging and the distribution of the 12 KIRIS forms are done so that each school has approximately equal numbers of each form of the test. Using this form spiraling process also aids with the equating process. If the spiraling process is poorly achieved a systematic equating error may result.

ITEM REPOONSE THEORY (IRT) MODEL

Lord (1953) made the important observation that observed examinee scores are not synonymous with ability scores: Ability scores are more fundamental because they are *test independent* whereas observed scores are *test dependent*. Examinees come to a test administration with fixed construct ability levels. These ability scores are test independent. However, examinee observed test scores are always dependent on the selection of items in that test.

Examinees will have lower observed scores on difficult tests and higher observed scores on easier tests, but their true ability scores remain invariant over any tests that measure the construct. Of course, over time, abilities may change because of instruction or other factors, but at the time of the test, each examinee will have a fixed ability score that is defined in relation to the construct. This ability remains independent over various samples of assessment items. Thus, a fairer estimate of student ability would be afforded using a method that provided true estimates of student ability regardless of the test items used in the assessment.

One solution lies in the concepts, models, and methods associated with item response theory (IRT, Lord, 1953, 1952). One of the more important reasons why KDE used unidimensional Item Response Theory (IRT) for KIRIS was that the IRT models permit examinees to be compared even though the examinees have not taken the same test. IRT models, using latent trait theory, provide a method to estimate an examinees latent ability. This single latent ability is measured on the same scale (theta) even though different examinees may have taken different forms of the same content area test. This test-free ability measurement is critical if multiple forms within and across years are used to obtain student ability levels.

All IRT models specify that an examinee's probability of answering a given question correctly depend on the examinee's ability and the characteristic of the item. This item characteristic is a function that relates the probability of student, of a given ability, getting an item correct. This relationship between item and student ability is called an items characteristic function or items characteristic curve (ICC). While it is possible to conceive of a number of IRT models, only a few models for open response items are in current wide spread use. KIRIS used Samejima's graded response model (Samejima, 1969). The Graded Response model assumes, in addition to the usual IRT assumptions that available categories to which an examinee responds can be ordered.

Examples would include a 5-point Likert rating scale or the 4-point rating scale for grading KIRIS open response items, or other scales representing levels of accomplishment or partial credit. Samejima extended the 2PL model (Lord, 1952, Birnbaum, 1968) into ordered categories. All initial student abilities using the graded response model were estimated using MULTILOG ® (1991).

ITEM CALIBRATION SAMPLES

IRT equating is usually a two-step process, 1) item parameters are estimated, and 2) student thetas (or parameter estimates) are then scaled back to the base year IRT scale. In order to estimate item parameters a large representative student sample must be used. Generally the larger this sample the more stable the item calibration (item parameter estimates) will be. Typically, the larger the calibration sample the smaller the standard errors of the estimated item parameters will become. Complete student data without known errors or omissions is preferred for the calibration sample.

A calibration sample was selected for each year/grade combination. Each sample included at least four-fifths (ranges from 80.0% to 95.3%) of the entire population of students receiving scores on the assessment. Because of the large sample sizes, it was not necessary to create a separate calibration sample for each test within a year/grade. Students were excluded from the calibration sample for the following ordered reasons:

- Student excluded from the assessment for approved reasons,
- Student did not respond to one or more items (these items were coded as “B” blank) on a test in any of the four content areas,
- Student obtained a total raw score of 0 on a test in any of the four content areas,
- Student record did not have a valid form number (i.e., 01-12; generally students who did not participate in the testing but were not excluded), and
- Student did not take the same form of the test in all assessed content areas.

Table 9-1 provides the number of students included and excluded from each calibration sample. The reason for exclusion is also provided. Students who would have been excluded for more than one of the five reasons listed above are assigned to the exclusion category that occurs first in that list. For example, a student who did not respond to 3 items on the mathematics test and received a total raw score of 0 on the science test would be counted as excluded from the calibration sample for not responding to one or more items.

Beginning in 1994, blank responses were assigned a score of B and totally incorrect responses were assigned the lowest score point 0. For both the purpose of computing total raw scores (bullet 3 above) and for computing ability estimates after calibration, scores of B were assigned a value of 0.

SAMPLE CHARACTERISTICS

Each calibration sample contains students who entirely completed 1 of 12 test forms across all content areas tested at that grade. For years 1997 and 1998, grades 4 and 8 saw a reduction in the number of students that had one or more blanks responses. These reductions may be partly attributed to the shift of two content areas from each of these grades to an adjacent grade. The large addition in zero scores for grade 11 in 1997 and 1998 can be attributed to a slight definitional shift in 0 item scoring. There was little net effect on calibration sample size because there were similar count reductions of students not responding to one or more items. Fairly consistent percentages of grade 11 students (82.9%, 81.7%, 80.0%, and 80.2%) across Cycle 3 were in the calibration sample.

ITEM CALIBRATION RESULTS

Item parameters were estimated with Samejima's (Samejima, 1969) two-parameter graded response model for ordered responses using MULTILOG®. For each five-category (0, 1, 2, 3, 4) open-response item a single discrimination (slope) parameter and four difficulty (item-category threshold) parameters were estimated.

MULTILOG® software has a default of 25 cycles to estimate the item parameters under marginal maximum likelihood (MML) estimation. If more MML cycles were necessary additional were used so that convergence was possible. The default convergence criterion is 0.001. For each test, the number of cycles required to estimate the item parameters and the maximum intercycle parameter change after the final cycle are presented in table 9-2. As can be seen in table 9-2 all content areas, across all grades approached the convergence criterion.

FORMS EQUATING

In many common-item nonequivalent groups design applications, each form's item ability estimates are estimated at the time the form is administered and scored. Unlike KIRIS these different forms are not administered at the same time. In this type situation each form's parameter was estimated separately with each forms theta scores having a mean of 0 and a standard deviation of 1. If each form's parameters were estimated separately the student theta scores on each of the forms was not considered equivalent.

TABLE 9-1
CALIBRATION SAMPLES

Year	Total Population	EXCLUDED					Included in Calibration Sample	% of Total Population in Calibration Sample
		Valid Exemption	One or More Blank Responses	Total Score 0	Non Valid Form	Multiple Form		
ELEMENTARY								
Grade 4								
1995	50321	864	2610	573	1	34	46,239	91.9
1996	48482	723	2201	541	0	31	44,986	92.8
1997	46085	839	1024	487	0	64	43,671	94.8
1998	47370	906	967	179	49	124	45,145	95.3
Grade 5								
1997	45859	395	1195	285	0	48	43,936	95.8
1998	47581	440	1418	511	111	94	45,007	94.6
MIDDLE								
Grade 7								
1997	47135	698	2365	600	7	64	43,401	92.1
1998	50124	857	3225	109	66	67	45,800	91.4
Grade 8								
1995	51974	755	4439	496	0	91	46,193	88.9
1996	51640	939	4435	831	0	33	45,402	87.9
1997	46948	684	2524	502	0	45	43,193	92.0
1998	49581	783	3697	848	41	105	44,107	89.0
HIGH								
Grade 11								
1995	42095	398	6231	556	0	18	34,892	82.9
1996	41924	546	6641	462	0	9	34,266	81.7
1997	41647	464	6004	1800	0	67	33,312	80.0
1998	41174	515	6202	1326	36	59	33,036	80.2

TABLE 9-2
NUMBER OF CYCLES AND MAXIMUM PARAMETER CHANGE OF FINAL CYCLE OF ITEM PARAMETER ESTIMATION BY YEAR

		Year			
Grade	Content	95	96	97	98
Elementary					
4	Reading	18 (0.00092)	14 (0.00099)	14 (0.00047)	14 (0.00065)
4	Science	19 (0.00064)	21 (0.00059)	17 (0.00056)	18 (0.00053)
4	Math	20 (0.00080)	22 (0.00077)	N/A	N/A
5	Math	N/A	N/A	17 (0.00069)	17 (0.00070)
4	Social Studies	19 (0.00051)	20 (0.00067)	N/A	N/A
5	Social Studies	N/A	N/A	19 (0.00068)	18 (0.00068)
Middle					
7	Reading	N/A	N/A	11 (0.00061)	16 (0.00078)
8	Reading	21 (0.00027)	23 (0.00070)	N/A	N/A
7	Science	N/A	N/A	21 (0.00073)	19 (0.00073)
8	Science	25 (0.00069)	25 (0.00073)	N/A	N/A
8	Math	25 (0.00152)	25 (0.00165)	17 (0.00077)	15 (0.00070)
8	Social Studies	25 (0.00193)	25 (0.00185)	25 (0.00129)	28 (0.00080)
High					
11	Reading	22 (0.00103)	25 (0.00076)	25 (0.00073)	27 (0.00083)
11	Science	25 (0.00132)	25 (0.00075)	18 (0.00075)	19 (0.00067)
11	Math	25 (0.00263)	25 (0.00168)	26 (0.00064)	31 (0.00082)
11	Social Studies	25 (0.00225)	25 (0.00191)	25 (0.00074)	22 (0.00075)

One method of putting each form’s parameter estimates, thetas, on the same scale (thus making the scores equivalent) is to use various linear transformation methods (methods are discussed under across-year forms equating). However, within year KIRIS testing was done at the same time. All 12 forms of each content area tested were administered and scored within the same time period. Hence, there was no need to estimate each form’s parameters at different times. Thus, the parameters for all 12 forms could be estimated together in a single MULTILOG® software run. All that is necessary was to code the non-common items on form the student did not take as “not reached” (see MULTILOG manual for complete details). If within-year form item parameter estimation is conducted in this manner the resulting estimates will be on the same scale. After this simultaneous forms parameter estimation procedure was conducted, in-year student thetas on any of the 12 forms could now be considered equivalent for the same student ability level.

ACROSS-YEAR TEST EQUATING

Reading, Mathematics, Science and Social Studies

INTRODUCTION

After the within year calibration of item parameters were completed, the across year equating procedures were conducted. This equating process was necessary to find the linear transformation necessary to place all current year item parameters and ability estimates (thetas) on the KIRIS 1993 base year scale. Once each student's theta score is on the 1993 base scale the student's performance level can be identified. Appendix F shows the grade and content area theta cut points that were used to identify Novice, Apprentice, Proficient, and Distinguished student performance levels. These student performance levels were used to provide each schools academic accountability score. Using equated students' thetas, each student's performance level across each year of testing had the same underling scale. Since student performance levels were used for school scores then by extension each schools accountably score was on the same underling scale. Since the KIRIS accountability system was dependent on school score growth from one year to the next it was extremely important that true growth was measured using same underlying scale. The year-to-year equating process accomplished this task.

Test Structure. As noted in the within-year forms equating section, each content area (Reading, Mathematics, Science, and Social Studies) had 12 forms of the test with each form containing a set of common items across the 12 forms. Also, in all four years of Cycle 3, there were two operational matrix sampled items administered in each of the 12 forms in each of the content areas.

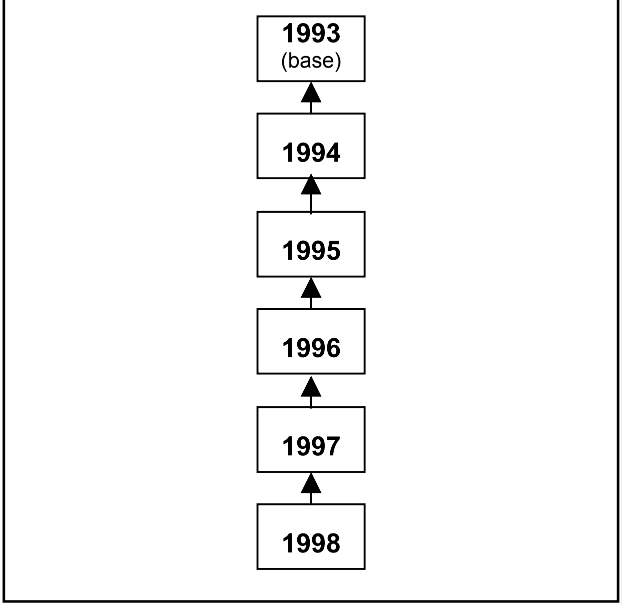
The matrix sampled items not only provided the intended additional content coverage but also provided a means to link each years testing back to 1993. Appendix G provides the Cycle 3 locations (form and item number) by content area of the open-response matrix items that were the same (common) in two adjacent years for each of the four years of Cycle 3. To linguistically differentiate between the common items across forms and the common items that link tests between years it has been the practice to identify the common items that link between year tests as linking items. The prior years linking item location information can be found in the Chapter 9 Appendix of the Cycle1 and Cycle 2 Technical Manual.

Equating Design. As can be seen by the test structure KIRIS has a number of linking (common) items between each adjacent year. Figure 9-1 provides a visual representation of the pattern of single link plan used by KIRIS. Tables 9-3a, 9-3b, and 9-3c indicate the initial number of linking items for each grade and content area.

There are several ways to transform or calibrate item parameter estimates from the current year's scale on to another using the linking items. Loyd and Hover (1980) uses a method known as *mean/mean*. Here the mean of the a-parameter and the means of the b-parameter estimates of the linking items are used to find the rescaled parameter estimates (Kolen and Brennan, 1995). Another method described by Marco (Marco, 1977) is the *mean/sigma* procedure. This method uses the standard deviations and means of the linking item b-parameters (Appendix H).

After the initial calibration of item parameters, *mean/sigma* equating procedures were conducted as described in the Cycle 3 equating plan (Appendix H). This was accomplished by first linking the appropriate grade and content area 1997 test to the grade and content area test given in 1996 so as to place the 1997 parameters on the 1993 metric. Then, using the 1997 adjusted parameters, the 1998 test (for that same grade and content area) was linked to the 1997 test and thus consequently to the base test year of 1993. [There was no need to equate the first two years of Cycle 3 (1995 and 1996) back to the 1993 scale because that equating process had been previously completed during Cycle 2.] Thus, the end result of a total of 48 equating processes for Cycle 3 (12 equating processes per year for four years) was to place all test parameters and theta values onto the same 1993 metric.

Figure 9-1
KIRIS Single Link
Equating Plan



EQUATING RESULTS

Equating results are summarized in tables 9-3a, 9-3b, & 9-3c. The final slopes and intercepts presented yield the transformations necessary to place item parameters and ability estimates for 1995, 1996, 1997, and 1998 onto the 1993 scale. Final slopes and intercepts were applied to the student ability estimates for all students receiving scores, not just those in the calibration sample. Distributions of scores in the theta ability estimate metric across the four years of the third accountability cycle are presented in table 9-4.

TABLE 9-3a
Equating Results -- Open Response
Elementary School

Grade	Subject	Year	Linking				Resulting Equating			
			Initial No. Of Items	Items Dropped	Total Points (b's) Dropped	Points[1] (b's) Left	Slope	Initial Intercept	Intercept Adjustment For Scorer Differences	Final Adjusted Intercept[2]
Grade 4										
4	Reading	1995	9	0	4	32	0.8970	0.8420		0.8420
		1996	11	0	3	41	0.8692	0.8720		0.8720
		1997	10	0	4	36	0.8316	1.0841		1.0841
		1998	12	1	3	41	0.8410	0.9266		0.9266
4	Science	1995	10	0	1	39	0.8290	0.6290		0.6290
		1996	12	0	8	40	0.7972	0.4135	0.1209	0.5344
		1997	12	0	16	32	0.7318	0.8499	-0.1070	0.7429
		1998	12	0	11	37	0.8413	0.6233	0.1617	0.7850
Grade 5										
5	Math	1995	10	0	8	32	0.770	0.640		0.640
		1996	12	0	1	47	0.7479	0.5951		0.5951
		1997	12	0	6	42	0.8349	0.9992		0.9992
		1998	13	0	0	52	0.8324	0.9934		0.9934
5	Soc. Studies	1995	9	0	2	34	1.0670	0.3630		0.3630
		1996	8	0	4	28	0.8735	0.3126	0.1304	0.3126
		1997	12	0	13	35	0.9103	0.6282		0.6282
		1998	12	1	10	34	0.9607	0.8536	-0.1731	0.6805
5	Arts & Hum.	1995	4	0	1	15				
		1996	5	0	0	20	1.0094	0.7528	-0.1517	0.6011
		1997	11	0	6	38	1.0801	1.1254		1.1254
		1998	12	0	4	44	1.0689	1.1351		1.1351
5	PL/VS	1995	6	0	7	17				
		1996	5	0	1	19	0.9421	0.7605		0.7605
		1997	10	0	3	37	1.0682	1.5148		1.5148
		1998	12	1	6	38	1.1554	1.2757		1.2757

1. When item score points (b's) are first removed from the equating in step 9E of the "Equating Plan Accountability Cycle 3" (Appendix H), the procedures call for the entire item to be removed unless the total number of b's is less than 40, in which case only the affected pair of b's (i.e., both the predecessor year b and the current year b) should be removed. In every case, removing entire items would have resulted in fewer than 40 b's remaining. Therefore, in all cases, only the affected pair of b's were removed.
2. Where applicable, scorer adjustment noted in a preceding column is included in the final Intercept figure.

TABLE 9-3b
Equating Results -- Open Response
Middle School

Grade	Subject	Year	Linking				Resulting Equating			
			Initial No. Of Items	Items Dropped	Total Points (b's) Dropped	Points[1] (b's) Left	Slope	Initial Intercept	Intercept Adjustment For Scorer Differences	Final Adjusted Intercept[2]
Grade 7										
7	Reading	1995	16	0	18	46	0.788	0.386		0.3860
		1996	11	0	5	39	0.7214	0.4156		0.4156
		1997	11	0	4	40	0.7845	0.3426		0.3426
		1998	11	0	7	37	0.8051	0.3681	-0.1119	0.2562
7	Science	1995	11	0	12	32	0.8110	0.2060		0.2060
		1996	8	0	2	30	0.7151	0.0031		0.0031
		1997	8	0	1	31	0.6711	0.0447	-0.0214	0.0233
		1998	12	2	5	35	0.7342	-0.0266		-0.0266
Grade 8										
8	Math	1995	10	0	3	37	0.819	0.537		0.5370
		1996	12	0	2	46	0.7509	0.5326		0.5326
		1997	13	0	3	49	0.8789	0.6029		0.6029
		1998	11	1	3	41	0.8298	0.5886		0.5886
8	Soc. Studies	1995	9	0	4	32	1.0390	0.4590		0.4590
		1996	11	0	8	36	0.8808	0.2424	0.0237	0.2661
		1997	11	0	8	36	0.8956	0.3184		0.3184
		1998	12	1	3	41	0.8873	0.3827	-0.2077	0.1750
8	Arts & Hum.	1995	6	0	9	15				
		1996	6	0	1	23	0.7495	0.3796	0.0852	0.4649
		1997	9	0	4	32	0.8893	0.7501	-0.0727	0.6774
		1998	12	1	1	43	0.9960	0.7972		0.7972
8	PL/VS	1995	6	0	12	12				
		1996	6	0	6	18	1.0828	0.4989	-0.1127	0.3862
		1997	14	0	6	50	1.1210	0.5587	-0.3097	0.2490
		1998	12	1	1	43	1.1817	0.5329		0.5329

[1] When item score points (b's) are first removed from the equating in step 9E of the "Equating Plan Accountability Cycle 3" (Appendix H), the procedures call for the entire item to be removed unless the total number of b's is less than 40, in which case only the affected pair of b's (i.e., both the predecessor year b and the current year b) should be removed. In every case, removing entire items would have resulted in fewer than 40 b's remaining. Therefore, in all cases, only the affected pair of b's were removed.

[2] Where applicable, scorer adjustment noted in a preceding column is included in the final Intercept figure.

TABLE 9-3c
Equating Results -- Open Response
High School

			Linking				Resulting Equating			
Grade	Subject	Year	Initial No. Of Items	Items Dropped	Total Points (b's) Dropped	Points[1] (b's) Left	Slope	Initial Intercept	Intercept Adjustment For Scorer Differences	Final Adjusted Intercept[2]
11	Reading	1995	15		25	35	0.9300	0.5590		0.5590
		1996	11	0	8	33	0.8736	0.5585		0.5585
		1997	10	0	8	32	0.9749	1.2345	-0.1176	1.1169
		1998	12	0	3	45	0.8903	1.1554		1.1554
11	Science	1995	17		28	40	0.924	0.6020		0.6020
		1996	9	0	4	32	0.8695	0.6075		0.6075
		1997	11	0	12	32	0.8543	0.8482		0.8482
		1998	12	0	9	39	0.7601	0.8262		0.8262
11	Math	1995	11	0	3	41	0.923	0.3810		0.381
		1996	9	0	2	34	0.8319	0.5186		0.5186
		1997	13	0	3	49	0.8043	0.7059		0.7059
		1998	12	0	3	45	0.8245	0.6220		0.6220
11	Soc. Studies	1995	15		17	43	1.0300	0.6070		0.6070
		1996	12	0	16	32	0.9694	0.1544	0.2859	0.4403
		1997	12	0	12	36	1.0434	0.7770		0.7770
		1998	12	0	4	44	1.0337	0.9330		0.9330
11	Arts & Hum.	1995	6		8	16				
		1996	7	0	7	21	0.8126	0.4912	-0.3070	0.1842
		1997	12	0	4	44	0.8351	0.2171		0.2171
		1998	12	0	2	46	0.9613	0.5056		0.5056
11	PL/VS	1995	6		9	15				
		1996	8	0	12	20	0.8876	1.0284	-0.4632	0.5652
		1997	13	0	9	43	1.0674	0.9147		0.9147
		1998	13	1	8	44	1.0554	1.0215		1.0215

1. When item score points (b's) are first removed from the equating in step 9E of the "Equating Plan Accountability Cycle 3" (Appendix H), the procedures call for the entire item to be removed unless the total number of b's is less than 40, in which case only the affected pair of b's (i.e., both the predecessor year b and the current year b) should be removed. In every case, removing entire items would have resulted in fewer than 40 b's remaining. Therefore, in all cases, only the affected pair of b's were removed.

2. Where applicable, scorer adjustment noted in a preceding column is included in the final Intercept figure.

**TABLE 9-4
DISTRIBUTION OF STUDENT THETA's**

		Mean (Standard Deviation), By Year			
Gr.	Content Area	1995	1996	1997	1998
Elementary					
4	Reading	0.777 (0.865)	0.811 (0.835)	1.055 (0.779)	0.897 (0.796)
4	Mathematics[1]	0.586 (0.727)	0.551 (0.695)		
4	Science	0.569 (0.769)	0.484 (0.731)	0.826 (0.652)	0.762 (0.734)
4	Social Studies[1]	0.283 (1.019)	0.249 (0.847)		
5	Mathematics			0.969 (0.752)	0.948 (0.784)
5	Social Studies			0.597 (0.825)	0.636 (0.881)
Middle					
7	Reading			0.283 (0.781)	0.184 (0.808)
7	Science			-0.002 (0.625)	-0.097 (0.711)
8	Reading[2]	0.290 (0.820)	0.318 (0.758)		
8	Mathematics	0.459 (0.790)	0.451 (0.727)	0.541 (0.823)	0.496 (0.796)
8	Science[2]	0.122 (0.786)	-0.079 (0.685)		
8	Social Studies	0.325 (1.063)	0.151 (0.883)	0.247 (0.874)	0.061 (0.882)
High					
11	Reading	0.416 (0.950)	0.404 (0.917)	1.019 (1.046)	0.954 (0.987)
11	Mathematics	0.417 (0.845)	0.389 (0.818)	0.552 (0.765)	0.459 (0.815)
11	Science	0.464 (0.922)	0.457 (0.864)	0.662 (0.860)	0.666 (0.772)
11	Social Studies	0.442 (1.037)	0.273 (0.970)	0.554 (1.079)	0.667 (1.151)

[1] Grade 5 in 1997 and 1998
[2] Grade 7 in 1997 and 1998

ARTS & HUMANITIES AND PRACTICAL LIVING/VOCATIONAL STUDIES

A modified version of the procedure used to equate reading, mathematics, science, and social studies was developed to link arts & humanities (A&H) and practical living/vocational studies (PL/VS) across and within years, due to the ways in which the A&H and PL/VS tests differed in their design from the other content areas. Beginning in 1995, a separate test section was created for A&H and PL/VS. All students completed two items in each content area, and items were totally matrix sampled and scored separately from reading, mathematics, science, and social studies.

Linking the arts & humanities and practical living/vocational studies tests across years was accomplished through a modification of the common item equating procedures used to equate the reading, mathematics, science and social studies open-response tests. The two modifications made to the original process were a change in the procedure for estimating the initial difficulties and a change in the cutoff for eliminating items from the equating step.

Initial item difficulties were estimated by computing a log-odds ratio based on the cumulative percentage of students scoring each point or above. For each item, four difficulties were computed corresponding to the percentage of students scoring 1 or above, 2 or above, 3 or above, or 4. This process has been used to estimate item difficulties for arts & humanities and practical living & vocational studies items since the initial year of testing in those content areas. An example is provided in Table 9-5.

**TABLE 9-5
EXAMPLE OF COMPUTING THE INITIAL ESTIMATES
OF A&H AND PL/VS ITEM DIFFICULTIES**

Scores	% scoring (P)	% not scoring (Q)	Difficulty: ln(Q/P)
1, 2, 3, or 4	90.6	09.4	-2.27
2, 3, or 4	47.0	53.0	0.12
3 or 4	14.0	86.0	1.82
4	02.8	97.2	3.55

Based on the differences in item difficulties from year to year, the cutoff for eliminating items and individual score points from equating was raised from a delta of 0.40 to 0.60 because of the limited number of items available across years. The lack of items available for linking Arts & Humanities and Practical Living & Vocational Studies tests was due in large part to the length of the tests. In 1995 each of the two tests contained 12 items and starting in 1996 each test contained 24 items. The number of linking items is provided in Appendix G and available linking points are provided in the Tables 9-3a, 9-3b, and 9-3c.

RESCORING ANALYSIS

As part of the equating process, a rescoring analysis was conducted on the linking items to evaluate if the scoring of these items had changed from the previous year to the current year of testing. In a criterion referenced scoring system, the scoring of an open response item should be the same no matter what year that item was evaluated.

The rescoring process started with the selection of 50 papers at each of the five score levels for each linking item. In 1998, to reduce the amount of time necessary to retrieve prior year tests for rescoring, all linking items across all subjects were rescored for each of the rescored papers. This procedure of course, will not yield exactly 50 students for each of the scoring levels. However, because oversampling was used the number of test items sampled did approach at least 50 students per scoring level. This procedure increased the speed in which the rescore sample was drawn and did not negatively affect sensitivity to any change item scoring.

The scores from this rescoring process are then compared to their original scores from the previous year. If it can be determined that a systematic scoring change occurred between the current and previous years scoring of these linking items for a grade and content area a statistical adjustment was made. If adjustment were made to the b parameter intercept they are documented on tables 9-3a, 9-3b, & 9-3c. Appendix I provides a more complete explanation of the analysis of the rescoring process.

GRADE SHIFT ADJUSTMENTS DUE TO CHANGE IN GRADES TESTED

For Reading, Mathematics, Social Studies and Science at grade levels, which remain constant between 1996 and 1997, the Cycle 3 equating plan was directly followed. For tests moved from grades 4 and 8 to grades 5 and 7, respectively, grade shift adjustments were necessary.

The Cycle III grade shift adjustments were determined, for grades 4 and 8, through the administration of existing tests at adjacent grades in spring of both 1996 and 1997. In 1996, grade 5 students along with accountable grade 4 students took grade 4 tests in mathematics, social studies, and arts & humanities and practical living/vocational studies at twenty-six selected schools. Also, grade 7 students along with accountable grade 8 students took grade 8 tests in reading and science at thirteen selected schools. In 1997 the reverse occurred, grade 4 students along with the now accountable grade 5 students took the mathematics, social studies, and arts & humanities and practical living/vocational studies KIRIS test at many of the same selected 1996 schools. Also, the grade 8 students along with the now accountable grade 7 students took the KIRIS reading and science test at many of the same selected 1996 schools. The original design intent was to use the difference between adjacent grade accountability index scores in schools participating in this study to serve as the Cycle 3 baseline adjustment. The performance standards for the adjacent grades were to be computed so that, on average, schools would have comparable accountability scores to those of the pervious accountable grade.

However, “there were clear differences between 1996 and 1997 results that indicated proficiency level standards established from the 1996 data did not adequately cross-validate in the 1997 data” (Wise, L., 1998). Appendix J provides a full summary of the grade shift adjustment study reviewed by HumRRO. A very plausible explanation for this finding could be the change in student motivation because of the nature of the testing situation. In one testing situation the students took the test in a “research only “ condition. All involved knew that these test scores were not part of the accountability system. In the second condition all persons taking and conducting these testing knew that the tests were accountable to the school and that the students would be receiving proficiency levels based on their performance. Thus, student motivation and possible teacher emphases on content areas associated with the newly accountable grades could have had a harmful impact on the study.

Because of these initial finding, a reanalysis was conducted combining 1996 and 1997 data using only schools that were involves in both years of the two studies. It was thought that combining these data would in effect balance the two differences in the collection process. Although both a linear and an equipercentile approach were used to analyze the data the more parsimonious equipercentile equating procedure was viewed as slightly better than the nearly identical linear analysis results. Using the equipercentile equating method, equal percentages of Novice, Apprentice, Proficient and Distinguished scores in the adjacent grades were accomplished.

The grade shift study produced Grade 5 theta cutpoints, which define performance levels that were slightly higher than for the same content areas that were previously tested in grade 4. These results were expected. Grade 7 theta cutpoints were slightly less than the cutpoints for the same content areas previously tested in grade 8. These results were also expected. Appendix F provides a listing of all theta cutpoints used in Cycle 3.